CORSO DI INTRODUZIONE AI BIG DATA

Primo Progetto

Si supponga di avere a disposizione un file di testo generato da un sistema di billing di una catena di supermercati che contiene, per ciascuno scontrino, una riga con la data e la lista dei prodotti acquistati, separati da una virgola.
Per esempio:

```
2015-03-21,uova,latte,pane,vino
2015-05-18,pesce,pane,insalata,formaggio
......
```

Il file puo essere creato autonomamente. E' disponibile un progetto Java chiamato "Data Generator" per la generazione automatica di un file con tale formato sul sito del corso.

Progettare e realizzare in: (a) MapReduce e (b) R:

1. Un job che sia in grado di generare, per ciascun mese del 2015, i cinque prodotti piu venduti seguiti dal numero complessivo di pezzi venduti. Per esempio:

```
2015-01: pane 852, latte 753, carne 544, vino 501, pesce 488
2015-02: latte 744, burro 655, uova 585, birra 498, pane 457
...
```

2. Un job che, dato un file in un formato a piacere contenente il costo di ciascun prodotto, sia in grado di generare, per ciascun prodotto, l'incasso totale per quel prodotto di ciascun mese del 2015. Per esempio:

```
pane 1/2015:12340 2/2015:8530 3/2015:9450 …
latte 1/2015:11987 2/2015:10980 3/2015:12350 …
...
```

3. Un job in grado di generare, per ciascuna coppia di prodotti (p1,p2): (i) la percentuale del numero complessivo di scontrini nei quali i due prodotti compaiono insieme (supporto della regola di associazione p1→p2) e (ii) la percentuale del numero di scontrini che contengono p1 nei quali compare anche p2 (confidenza della regola di associazione p1→p2)

```
pane,latte,30%, 4%
vino,uova,23%, 4%
latte,pane, 30%, 7%
...
```

Per ciascun job bisogna illustrare e documentare in un rapporto finale:

- Una possibile implementazione MapReduce (pseudocodice) e R (pseudocodice)
- Tabella e grafici che confrontano i tempi di esecuzione in locale dei vari job con dimensioni variabili dell'input e del numero di nodi su cluster
- Il relativo codice completo MapReduce e R (da allegare al documento)
- Un test di uso con file di input (di piccole dimensioni) e file di output (da allegare)

Tutte le specifiche non definite in questo documento possono essere scelte liberamente. Consegnare il rapporto prima dell'esame via email.